

# **Some Considerations for Designing Summative Assessment Components With Item Types Beyond Machine-Scorable Item Types**

Brian Gong, Center for Assessment<sup>i</sup>  
April 16, 2010

The purpose of this document is to help consortia intending to respond to the RTTT Assessment NIA. The document frames key design choices that a consortium will need to make. As the consortium makes decisions regarding the design choices presented in this document, the consortium will be generating the conceptual design of its summative assessment component in response to the NIA requirements.

## **General Orientation**

The Comprehensive Assessment System Notice of Intent to Award (NIA) has the primary focus of states helping more students graduate from high school better prepared for college. The NIA assumes that some key ways states can achieve this goal involve content standards, assessments, and uses of assessment. The key uses include school accountability according to ESEA, educator effectiveness evaluation, and providing assessment information to local educational agencies, educators, and others.

The NIA challenges consortia to go beyond current assessment practice, which has focused on standards, assessment, and accountability. Some of the modifications specified in the NIA include:

- Content standards anchored on college-readiness
- Common achievement standards
- Summative assessments that assess things that were previously not assessed because they were technically difficult or not focused upon
- Attention beyond summative assessment to other assessment processes and types (e.g., interim and formative)
- [etc.]

## **General Criteria for Assessment Design**

The NIA specifies three criteria for the assessment design:

- Innovation
- Feasibility
- Consistency with the consortium's theory of action [p. 33]

**1. What assessment information is needed for the consortium's theory of action? (See p. 6, section on "Developing a Theory of Action.")**

**2. What relative balance between “innovation” and “feasibility” will the consortium take? What does that look like in terms of assessment? (For innovation especially, see p. 6, section on “Developing a Theory of Action.”)**

### **Assessment Design NIA Considerations**

The NIA states, “In determining the extent to which the design has these attributes [innovation, feasibility, consistency with the theory of action], we will consider”—and provides a set of criteria. (See Appendix for full set of criteria for Assessment Design.)

Two criteria seem to imply the desirability of using performance assessments along with other types of item types:

(b)(i) How the assessment system will measure student knowledge and skills against the full range of the college- and career-ready standards, including the standards against which student achievement has traditionally been difficult to measure...

(c)(iv) The number and types of items (*e.g.*, performance tasks, selected responses, brief or extended constructed responses) and the distribution of item types within the component, including the extent to which the items will be varied and elicit complex student demonstrations or applications of knowledge and skills (descriptions should include a concrete example of each item type proposed); and the rationale for using these item types and their distributions;

### **Design Choices for Performance Assessments**

The following questions pose choices for the consortium. As the consortium makes decisions about these choices, it will be specifying the design for performance assessments and their use. The use of this document will help promote the coherence and completeness of the design specifications.

This document focuses on a summative assessment for making student-level determinations of “college-readiness” that would then be aggregated and used for school accountability in an NCLB-type model. Other uses will require modifications to the specifics in the section below.

- 3. What purposes would you like performance assessments to play?**
- a. Performance assessments increase validity of what is measured, i.e., the construct would not be measured adequately without the performance assessment**
  - b. Performance assessments increase the utility of assessment data, i.e., the assessment reports/data are more useful because of inclusion of performance assessments—e.g., more reliable, provide more**

**information.**

- c. Performance assessments influence learning and teaching by providing information other than the assessment results.**

**4. How will you determine where performance assessments are conceptually most appropriate?**

- a. For which content/performance standards are performance assessments needed to adequately measure the construct? (See 3.a.)**
  - i. Are performance assessments necessary to adequately assess a complete standard?**
  - ii. Are performance assessments necessary to adequately assess part of the content standard?**
  - iii. Are performance assessments necessary to adequately assess combinations of content standards or clusters of standards related to the overall definition of “college readiness”; e.g., college-readiness involves performance that integrates or otherwise combines multiple content standards? (See p. 9, discussion of “College-readiness Standards and Growth.”)**
  - iv. Are performance assessments necessary to adequately provide evidence for attainment of the achievement level? (See discussion below.)**
  - v. How will performance assessments that are necessary to adequately measure the construct be developed so they are fair to all students, including students with disabilities, etc.?**
  - vi. How is growth defined in the construct and how do performance assessments help measure growth? (See p. 9, discussion of “College-readiness Standards and Growth.”)**
  - vii. How is information from performance assessments related to constructs relevant to intended uses other than assessing student college-readiness, such as school accountability, educator effectiveness evaluation, program evaluation, etc.?**
- b. How will information from performance assessments increase the utility of reports and support better uses? (See 3.b.)**
  - i. Will the results from performance assessments be reported separately from the results from other assessment pieces in ways that inform useful action? If so, how?**
  - ii. Will the results from performance assessments be combined with other assessment data in a way that the results can be attributed to the performance assessments, e.g., “to score Advanced, the student had to score well on the performance assessment”?**
  - iii. Will the performance assessments make the score more reliable or otherwise enhance the technical properties of the assessment interpretation, other than increasing validity (validity is addressed in 3.a. and 4.a., and consequential validity is addressed in 3.c. and 4.c.)? If so, how?**
- c. How will performance assessments influence learning and teaching other than by providing assessment results information?**
  - i. Released performance assessments provide clearer information about desired learning and assessment targets that will be emulated in instruction or that teachers will work towards.**

- ii. Participation in the performance assessments is a valuable learning experience for students.
  - iii. Etc.
- 5. Place of performance assessments in design of assessment “form”**
- a. What is sufficient amount of information needed to make the desired judgment, e.g., how much evidence is needed to declare a student is college-ready?
    - i. The amounts may be set ahead of time for all students, or other models may be used, such as: CAT typically optimizes on reaching an estimate of student ability with a pre-established degree of precision (reliability); a multi-stage model of assessment might give a common “locator” assessment to all students and then a different “Stage 2” assessment based on student performance on the locator—the design criterion might be to maximize content coverage or assessment of student transfer for high-end students but assessment of areas of weakness for low-end students, etc.
  - b. What amount of information from performance assessments is needed?
  - c. How varied must the performance assessment tasks be from year to year?
  - d. Etc.
- 6. Operational considerations to inform design of performance assessments**
- a. Must performance assessments be administered on computer?
    - i. Prompt and any supporting information presented on computer?
    - ii. Response entered on computer? How? Technology available? Technology available for particular performance assessment task’s demands?
      - 1. “Gridded” type responses
      - 2. “Drag” or “click-to-identify” type responses to items with highly structured possible responses
      - 3. Special interface conventions, e.g., equation editor (the response is like what a student would write without the computer)
      - 4. Special software functions, e.g., CAD application (the response is shaped heavily by the software)
      - 5. etc.
    - iii. Scored by computer?
    - iv. Report generated by computer?
  - b. Must performance assessment be administered by humans?
  - c. How much time is allowed for administration of performance assessments?
  - d. What are the assessment development implications, e.g., cost, time, available expertise? (See section on Assessment Development)
    - i. How many performance assessments must be developed new per administration?
    - ii. How much does it cost to develop the required performance assessments?
  - e. Scoring

- f. Reporting**
  - i. What information from performance assessments will be reported, how? (See 4.b.)**
  - ii. When must which information from performance assessments be reported?**
- g. Interpretation and use**
- h. Additional requirements for growth**
- i. Additional requirements for other uses (e.g., teacher effectiveness evaluation, program evaluation, etc.)**

## Developing a “Theory of Action” for the Assessment System

The consortium must include a “theory of action” in its proposal. A “theory of action” is an argument of what it proposes to do in order to achieve its stated goals.<sup>1</sup> The argument includes attention to how achievement of one part is connected to other parts, and ultimately to the attainment of the goal.

A strong theory of action includes these parts:

1. A compelling **statement of the problem** – What problem is worth working on? Is it the right problem?
2. An insightful **statement of the solution** – How can the problem be solved? The solution should be *innovative* in that it embodies how this solution will work when previous solutions did not.
3. A suitable **elaboration of the solution** – Explain the main elements of the solution and how they are related. The elements should be given in enough detail to encompass the true power of the solution. Particular attention should be paid to innovative aspects and parts that may be problematic. Contingencies are often appropriate, identifying likely deviations from the solution plan and how they will be handled.
4. Inclusion of **assessment data and uses** – Since this is a proposal for development of a comprehensive assessment system, the theory of action must clearly show how the assessment system will be part of the solution. In particular, the elaboration of the solution should include what assessment data will be produced and how it will be used to help achieve the goals.
5. A theory of action typically should be linked with a **research and evaluation plan** that will produce **evidence that the theory of action is effective**, and also valuable formative evaluation information to help **improve the theory and its implementation**.

One common challenge to developing theories of action that center on assessment and accountability systems is that the linkages between steps has not been specified. This has resulted often in difficult problems being glossed over (“and then a miracle occurs”<sup>2</sup>) or not being able to monitor and improve the theory of action. There have been some philosophical reasons for not fully specifying a theory of action’s linkages of “how” the goal will be achieved. Notably, modern assessment and accountability models were heavily influenced by outcomes-based measurement and management by objective models that stipulated effective management provided clear goals, timely and accurate feedback, and strong motivation (e.g., rewards, supports, and sanctions if necessary), but left up to those responsible for actual implementation how to achieve the goals. Thus, many accountability models—including NCLB—touted it as a feature that there was no specified “how the goal was to be achieved.” That was not only to be left up to local

---

<sup>1</sup> I have presented an expanded definition of “theory of action” than is typically found in the academic literature around assessment and accountability systems. I have recast “theory of action” as an argument whereas many other authors emphasize the explication of causal connections.

<sup>2</sup> That is, the result has rarely happened before and there is little plausible reason why it would happen now.

control, the model says, but also that it is so varied that it is impossible to describe let alone prescribe.

A second, related challenge to filling out the linkages in a theory of action of this-is-used-to-make-this-happen-which-then-is-used-to-make-this-happen-etc. is that the actions often crossed boundaries of the educational system. In particular, there are many challenges in a state specifying what districts *will* do, and similarly for districts to specify what schools *will* do, and for schools to specify what a teacher *will* do, and for a teacher to specify what a student *will* do.<sup>3</sup> The state clearly should be responsible for elaborating its theory of action. A stronger, more effective state theory of action will include connections with other levels of the system—connections to the districts’ theories of action, and connections with the larger system components that affect education, such as higher education opportunities, economic development, early childhood supports, teacher and principal supply and distribution throughout the state, and so on.

### An Example Theory of Action as an Argument

A disclaimer: State leaders can make an argument much better than this example. In fact, I am embarrassed about its shortcomings. I hope you don’t focus on this example’s weaknesses, but rather see that a “theory of action” argument is something that you do all the time. The needs for the proposal are an emphasis on assessment and the right level of detail—and, of course, great problems, solutions, elaborations, and means to evaluate and improve coherently related together.

*Statement of Problem/Goal:* Our State has done well. But what has been good enough up to now will not be good enough in the future. There are three critical needs we have less than a generation to address. First, we need more students to graduate from high school ready for college, and to enter college. Currently about 40% of the students in our state who enter high school graduate and enroll in college. But about 4 in 10 of these students arrive at college and find they are not prepared to take credit-bearing courses. But our economy is changing. We need to increase our percent of students graduating ready for college dramatically. We have set the goal of 60% by 2022 (when today’s kindergarteners will graduate) and reduce the remediation rate from 40% to 10%, and 80% college-going in twenty-five years. Second, we need to increase the proportion of students who are strongly prepared for college, and especially those who major in science, technology, mathematics, and engineering. Currently about 8% of the students in our state enroll in college prepared for a major in those areas. We propose to double that by 2022. Third, about 25% of the students who start high school drop out before completing. In many ways this is the most tragic and difficult problem to solve, but the one that calls us most clearly. We propose that we slash the cumulative dropout rate by half to 12% by 2022. These are not new problems. But we have a sharper understanding of the nature of the problem, and we understand now that the past one-size-fits-all-solution is not what need for the future.

---

<sup>3</sup> There is also a philosophical argument about agency and whether one entity can or should “control” the outcomes enough to be held responsible for what another entity does.

*Statement of Solution (with emphasis on assessment):* We have three different challenges, and three different solutions. To reduce the dropout rate, the solution lies primarily in stronger schools with better teachers—85% of the students who drop out come from 20% of our schools, with known structural weakness in educator quality. We will help these districts and schools get *early intervention* and *strong feedback* on what works and what doesn't. We will do three things differently with our assessment that will help reduce dropouts. First, our school accountability system will distinguish the truly low from those that appear low but are actually growing. Second, the state will help LEAs implement periodic assessments that will provide feedback for individual students, teachers, and school programs several times per year so that instructional action need not wait for the state's end of year test. Third, we are developing an educator effectiveness evaluation program that will help teachers and building leaders of these students to identify their own strengths and weaknesses, and improve their weaknesses so that the most needy students have highly qualified teachers and leaders. And those educators who are persistently ineffective will not have their teaching certificates renewed.

For schools in the middle, the key is implementing strong curricula that adequately prepare students for college. We have identified and *corrected gaps* in our current content standards, and will *assess* students against our improved *college-ready content and performance standards*. In addition, we realized that college-readiness involves more than just the academic knowledge and skills that can be measured in a two-hour test. We are implementing an *expanded assessment* that will address the critical college-ready aspects of integration of knowledge, student application and performance on complex tasks, and student choice so students develop an area of expertise that prepares them to choose a major.

Etc.

*Elaboration of the Solution (focus on assessment):* ...To reduce the college remediation rate from the current 40% to 10% by 2022, we will focus on opportunity and quality. By opportunity we mean the students who graduate from college but have not taken a curriculum that prepares them adequately for college. Our studies show this is 60% of our graduating students. We have adjusted our school accountability system to ensure that *middle schools* prepare students by... The other major problem is students who do take the recommended high school courses and get high grades but who do not pass the colleges' placement tests. We have instituted state end-of-course exams... These exams will be used with other evidence to certify that students knew the mathematics or English content of those required courses. We will use these tests to help districts select students for focused diagnosis with other diagnostic tests, and appropriate remediation. More importantly, we will monitor districts' use of the data to improve their programs; the pass rate discrepancy will be part of school accountability.

Etc.

*Inclusion of assessment data and uses:*

*Research and evaluation plan:*

## **Defining “College-Ready” and “Growth”**

### College-ready Standards: What is the right unit of analysis?

The NIA calls for states to address a problem long-discussed: the definition of student proficiency needed to be anchored in the real world. That is, there is a fundamental problem when substantial proportions of high school students are declared “proficient” by the state test but are evidently not proficient according to other evidence, such as college remediation rates or job qualification assessments.

The NIA requires a consortium to adopt college-readiness content standards in English and mathematics. And the NIA requires the consortium to agree to set “college-readiness” achievement standards using empirical data so that human judgment of “college readiness” is checked with other data.

So content standards that identify what a students needs to know and be able to do to be “college-ready” are an essential element in the consortium’s theory of action about assessment and the interpretation and uses of assessment data. A consortium must be sure that the content standards for college-readiness are a) reasonably complete, and b) the assessments are specified at the right unit of analysis in the standards.

For whatever reasons, the focus in RTTT and by states has been on content standards of academic knowledge, so this comment is included for the sake of completeness. If the ultimate criterion is empirical success in college, then more is required than academic knowledge. The consortium should consider sets of standards that capture what is needed for college readiness/success. For example, one researcher has identified four types of knowledge and skills: a) academic knowledge, b) strategic knowledge and habits of mind such as problem solving, research skills, and appropriate attention to detail, c) social and non-cognitive skills, such as appropriate cooperation and self-discipline, and d) knowledge of the culture of college, such as how to apply for college and how learning in college is much less structured than attending high school.

The unit of analysis in regards to standards is of central importance to the consideration of performance assessments. Recent standards-based work has treated the individual standard as the unit of analysis for assessment. Thus in most testing programs an assessment item is developed and coded to a single standard or part of a standard. Popular alignment approaches count the correspondence between individual standards or parts of standards and individual assessment items. The main point is that most standards documents have fairly discrete statements of content as standards, and then may have additional statements about more specific content; that is, the standards are typically decomposed into a smaller unit, sometimes called an objective, a bullet, etc. However, proficiency or expertise in real-world situations such as college or work requires a student be able to integrate the knowledge and skills across standards, and usually across strands and often across domains (e.g., merge science and math, or reading and math).

Statements of proficiency or college-readiness should reflect these combinations of standards. And assessments should be designed to assess performance on combinations of standards larger than individual standards or parts of standards. Test blueprints should reflect such more direct assessment of such larger units of knowledge and skills. Performance assessments typically are better at assessing these larger units or combinations of standards than other assessment formats or “smaller” items.

### Defining Growth

Growth must be defined in order for assessments to be designed to measure growth. There are at least five conceptions of growth based in increase of knowledge or “content.” There are some other conceptions of growth that are based less directly on conceptions of content. The key issue that is often ignored is not the definition of growth or the measurement of growth, but the educational goal—what type of growth will be promoted in the schools as an appropriate learning goal. Once that is defined, then assessments of growth can be designed accordingly.

1. Growth is increase in performance on the same thing, towards mastery. Growth of this type is possible as long as the student did not have a perfect score, all the time. An instructional strategy to “help a student grow” under this conception of growth would be to help a student who scored proficient but did not get all the items correct to continue to study until s/he could get all [or more of] the items correct. An assessment designed to measure growth this way would consist of a carefully selected set of items where “all items correct” represented the top end of desired mastery, and increasing items correct represented increasing mastery. Note that we assume in our current typical assessments that more items correct = more proficient, regardless of which items are answered correctly or incorrectly. Many researchers and educators now propose that the items should be selected or developed to represent a progression where the order, and therefore which items are answered correctly, matters. The order is based on some definition of conceptual proficiency rather than merely item difficulty, since difficulty may be influenced by many things besides the desired learning progression.
2. Growth is learning one topic and then learning a more advanced topic in a sequence of content. Growth of this type is possible as long as there is more advanced content. An instructional strategy to “help a student grow” under this conception of growth would be to help a student who “mastered” the grade 4 content in math to go on to the grade 5 content. An assessment designed to measure this conception of growth would consist of assessment items carefully aligned to grade level sequences, without limits of which grade (or where within a grade sequence) a student might be assessed; in particular, it would expect an advanced student enrolled in a certain grade to be assessed on the content of the next grade, or a less advanced student to be assessed on the content of a lower grade than the one in which the student was enrolled.
3. Growth is increase in expertise on the same thing, where expertise incorporates a more powerful mental model, increased fluency, greater independence, or other

attributes of expertise. Growth in this conception is marked less by going on to new topics as it is by the depths and quality of understanding and ability to perform. An instructional strategy to “help a student grow” under this conception of growth would be to help the student move to an understanding of the deep structure of the discipline, understand constraints, limitations, and boundary conditions, develop fluency and automaticity, understand the reasons for why things are done as they are and what can be changed without affecting the essential, understanding multiple representations of the same content, etc. An assessment designed to measure this conception of growth would consist of assessment items designed for these aspects of expertise even though the “content” might appear to “stay the same.”

4. Growth is increase in integration across content and skills. Growth is the ability to not only solve the problems on the math worksheet that are all concerned with the same concept and the same general procedural approach, but to solve problems that require combining that knowledge and skill with other knowledge and skills previously learned. An instructional strategy to “help a student grow” under this conception of growth would be to circle back to previously learned knowledge and skills and help the student combine and apply them with the current knowledge and skills. An assessment designed to measure this conception of growth would consist of sequences of combinations of knowledge and skills, usually reaching back to previously learned content.
5. Growth is increase of knowledge and skills outside the sequence defined by the classroom. An instructional strategy to “help a student grow” under this conception of growth would be to allow the student, once s/he had completed the class assignments satisfactorily, to choose to do something else productive not related to the class curriculum, for example, to read a book of her/his own choosing once finished with the math assignment. An assessment of this type of growth would need to be quite individual.

To repeat, the consortium should define what is the conception of growth (may be combinations of the above) that schools will be asked to promote in their curriculum and daily instructional decisions. The state assessments of growth should be designed to match the learning model.

The term, “growth” when used with assessment now also commonly refers to some measurement interpretations that are often divorced for the above conceptions of growth, or at least do not specify the growth in content knowledge or skills.

6. Growth is increase in scores on a test that is constructed so that scores can be compared, e.g., a “vertical developmental scale.” The problem is that most scales do not document which type of “growth,” as discussed above, is being measured. Or to put it another way, it often is not clear whether a type of growth above would be reflected accurately by the change in scores.

7. Growth is an increase in the relative position compared to other students. This type of measurement can be very informative, but it is abstracted away from *what* the growth consists of, in terms of what the student knows or can do, as described above.

## APPENDIX

### Assessment System Design Criteria [pp. 33-34]

#### (A)(3) Assessment System Design (up to 55 points)

The extent to which the design of the eligible applicant's proposed assessment system is innovative, feasible, and consistent with the theory of action. In determining the extent to which the design has these attributes, we will consider—

- (a) The number and types of components (*e.g.*, through-course summative assessments (as defined in the NIA), end-of-year summative assessments, formative assessments, interim assessments in mathematics and in English language arts in the assessment system);
- (b) For the assessment system as a whole—
  - (i) How the assessment system will measure student knowledge and skills against the full range of the college- and career-ready standards, including the standards against which student achievement has traditionally been difficult to measure; and provide an accurate measure of student achievement, including for high- and low-performing students, and an accurate measure of student growth over a full academic year or course;
  - (ii) How the assessment system will produce the required student performance data (*i.e.*, student achievement data and student growth data (both as defined in the NIA) that can be used to determine whether individual students are college- and career-ready (as defined in the NIA) or on track to being college- and career-ready (as defined in the NIA);
  - (iii) How the assessment system will be accessible to all students, including English learners and students with disabilities, and include appropriate accommodations (as defined in the NIA) for students with disabilities and English learners; and
  - (iv) How and when during the academic year different types of student data will be available to inform and guide instruction, interventions, and professional development; and
- (c) For each component in mathematics and in English language arts in the assessment system--
  - (i) The types of data produced by the component, including student achievement data (as defined in the NIA), student growth data (as defined in the NIA), and other data;
  - (ii) The uses of the data produced by the component, including determining whether individual students are college- and career-ready (as defined in the NIA) or on track to being college- and career-ready (as defined in the NIA); informing determinations of school effectiveness for the purposes of accountability under Title I of the ESEA; informing determinations of individual principal and teacher

effectiveness for the purposes of evaluation; informing determinations of principal and teacher professional development and support needs; informing teaching, learning, and program improvement; and other uses;

(iii) The frequency and timing of administration of the component, and the rationale for these;

(iv) The number and types of items (*e.g.*, performance tasks, selected responses, brief or extended constructed responses) and the distribution of item types within the component, including the extent to which the items will be varied and elicit complex student demonstrations or applications of knowledge and skills (descriptions should include a concrete example of each item type proposed); and the rationale for using these item types and their distributions;

(v) The component's administration mode (*e.g.*, paper-and-pencil, computer-based, or other electronic device), and the rationale for the mode;

(vi) The methods for scoring student performance on the component, the estimated turnaround times for scoring, and the rationale for these; and

(vii) The reports produced based on the component, and for each report, its intended use, target audience (*e.g.*, students, parents, teachers, administrators, policymakers), and the key data it presents.

---

<sup>i</sup> This paper is one of a series of brief documents produced by the Center for Assessment designed to assist state consortium leaders in thinking about and preparing their responses to the United States Department of Education Notice Inviting Applications for Comprehensive and High School Assessment System grants. Funding has been provided by the Bill & Melinda Gates Foundation. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Gates Foundation.